Applied Energy 259 (2020) 114193

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system

Zhe Dong[a,b,*], Xiaojin Huang[a], Yujie Dong[a], Zuoyi Zhang[a]

[a] Institute of Nuclear and New Energy Technology (INET), Collaborative Innovation Center of Advanced Nuclear Energy Technology of China, Key Laboratory of Advanced Reactor Engineering and Safety of Ministry of Education, Tsinghua University, Beijing 100084, China
[b] State Key Laboratory of Nuclear Power Safety Monitoring Technology and Equipment, Shenzhen, Guangdong, 518172, China

HIGHLIGHTS

- Reinforcement learning is introduced to energy system dynamic optimization.
- A novel multilayer perception based reinforcement learning control is proposed.
- This new algorithm approximates the optimal control online without system model.
- Application to the optimized thermal power control of a nuclear reactor is given.
- Simulation results show the feasibility and satisfactory performance.

ARTICLE INFO

Keywords:
Energy system optimization
Reinforcement learning control
Neural network

ABSTRACT

Energy system optimization is important in strengthening stability, reliability and economy, which is usually given by static linear or nonlinear programming. However, the challenge faced in real-life currently is how to give the optimization by taking naturally existed energy system dynamics into account. To face this challenge, a multi-layer perception (MLP) based reinforcement learning control (RLC) method is proposed for the nonlinear dissipative system coupled by an arbitrary energy system and its local controllers, which can be able to optimize a given performance index dynamically and effectively without the accurate knowledge of system dynamics. This MLP-based RLC is composed of a MLP-based state-observer and an approximated optimal controller. The MLP-based state-observer is given for identification, which converges to a bounded neighborhood of the system dynamics asymptotically. The approximated optimal controller is determined by solving an algebraic Riccati equation with parameters given by the MLP-based state-observer. Based on Lyapunov direct method, it is further proven that the closed-loop is uniformly ultimately bounded stable. Finally, this newly-built MLP-based RLC is applied to the supervisory optimization of thermal power response for a nuclear steam supply system, and simulation results show not only the satisfactory performance but also the influences from the controller parameters to closed-loop responses.

1. Introduction

Energy system optimization such as improving the efficiency and enhancing the stability is crucial for not only the economic competitiveness but also the sustainable development of human society. Usually, the optimization is performed by solving a static linear or nonlinear programming problem. For example, the thermal efficiency can be improved based on exergy analysis that may be helpful to identify the sources of irreversibility [1]. However, the main challenge of energy system optimization faced in real-life is how to efficiently control the naturally dynamic energy systems so as to minimize or maximize a given performance index related to efficiency, stability, robustness and etc. The current results in the dynamical optimization of energy systems are mainly based on either classical optimal control theory or model predictive control (MPC) method.

Classical optimal control theory based on Pontryagin's maximum principle (PMP) and Bellman's dynamic programing (DP) principle is a mature discipline in designing optimal controllers for dynamical systems with specific cost capturing control objectives. PMP provides a necessary condition of optimality, while DP gives a sufficient condition

* Corresponding author.
E-mail address: dongzhe@mail.tsinghua.edu.cn (Z. Dong).

https://doi.org/10.1016/j.apenergy.2019.114193
Received 16 September 2019; Received in revised form 14 November 2019; Accepted 16 November 2019
Available online 26 November 2019

of optimality by solving a partial differential equation called Hamilton-Jacobi-Bellman equation. In [2], an optimal control is designed for the energy management of automotive power systems with battery or supercapacitor, which gives the most appropriate manner of electricity generation and storage so that the overall energy consumption and eventually the pollutant emissions can be minimized for a driving cycle. In [3], a supervisory control based on PMP is proposed to optimally allocate the demands of a group of n boilers in parallel, where the thermal power setpoint of every boiler is continuously optimized so as to minimize a combined cost cumulated in time and taking into account the dynamics of all individual boilers. In [4], an H_2-optimal regulator using both the supply-side and demand-side resources is given for grid balance in the presence of deep penetration of intermittent renewables. This optimal control can minimize the loss of total economic surplus induced by the deviations of frequency and voltage, and outperforms the conventional control policy by providing both faster response and lower cost in using reserve units for regulation services. Moreover, PMP has also been widely applied for the optimal control design of electric vehicles. A real-time optimal control strategy based on PMP is given to minimize the energy consumption of the electrical vehicles with the fuel cell and supercapacitor as the two power sources, where a Markov chain is adopted to predict the future power demand during a driving cycle [5]. An optimal energy management algorithm based on PMP and piece-wise approximated model is proposed for parallel plug-in hybrid electric vehicles (HEVs) [6]. An adaptive supervisory control is given in [7] based on PMP for on-line energy management optimization of a plug-in HEV, where the adaptation of the control parameters is driven by the feedback from state of charge (SOC). To further improve the operating efficiency of dual-motor-driven electric buses, a simple and robust power-management strategy was proposed in [8], where PMP is applied to optimize the control under three different driving cycles. Although optimal control laws based PMP or DP are effectively in minimizing the costs or consumptions, the corresponding design procedure is offline and requires complete knowledge of the system dynamics, which cannot be able to cope with dynamical uncertainties.

The model predictive control (MPC) method determines the current control input at each time step by solving an optimization problem for a finite future, where a dynamic model is used to predict the system behavior in response to a sequence of control inputs over a specific horizon so as to give the control decision at the current time step. This dynamic model is called prediction model which can be given by physics, measured input-output response as well as regression from operation data. Recently and currently, MPC is a hot spot in dealing with dynamical optimization for energy systems, especially for thermal systems, buildings, microgrids vehicles and nuclear plants:

(1) MPC of Thermal Systems

For the MPC of thermal systems, there have already been some promising results. Motivated by the fact pointed out in [1] that exergy destruction minimization at steady conditions cannot lead to significant reductions of energy consumption in dynamical operation of thermal systems, a MPC for the canonical four-component vapor compression system (VCS) was given in [9], which uses an exergy-based objective function to optimize the control actions for the VCS so as to maximize exergetic efficiency while achieving a desired cooling capacity. To address the uncertainty of heat transfer coefficients caused by the fouling of heat exchangers, a robust MPC with integral action was proposed for the shell-and-tube type heat exchangers of an industrial heat-exchanger network, which is capable to handle the uncertainties while assuring offset-free responses for process variables [10]. To effectively control the ice storage of a district cooling system, an MPC coupled by an ANN and a genetic algorithm (GA) was proposed in [11], where the ANN offers a fast prediction of ice storage and the GA is adopted for control action optimization. In [12], a MPC-based coordinated controller is given for the coupling system of a coal fired power plant (CFPP) and a solvent-based post-combustion CO_2 capture (PCC) process, which gives a flexible trade-off between the power generation and CO_2 reduction.

(2) MPC of Buildings

MPC methods have also been deeply studied for optimizing the operation efficiency of building systems. Motivated by the importance of energy management method for the buildings powered by intermittent renewables, an MPC using artificial neural networks (ANNs) was designed and implemented in a residential building [13], which provides a satisfactory prediction of the energy consumption of building and can further optimize the corresponding demand flexibility. In [14], a MPC with the prediction model given by deep time delay neural networks (TDNN) and regression trees (RT) is designed to enhance the energy efficiency of building systems. In [15], a novel MPC-based load aggregation method is given for the electricity distribution of a large cluster of buildings. In [16], a MPC is proposed using the weather and electricity cost predictions so as to improve the thermal efficiency of buildings with latent heat energy storage (LHES). To improve the energy efficiency of buildings, a MPC with the prediction model identified by the influence from the factors of building design, model structure, model order, data set, data quality, identification algorithm and initial guesses as well as software tool-chain to the variations in heating, ventilating and air conditioning (HVAC) is implemented [17]. In [18], a specialized MPC strategy is given to optimally manage the centralized HVAC system and the storage devices of buildings under thermal comfort and technological constraints. In [19], a MPC approach, which has the features of shrunk prediction horizon, self-correction and simple parameter determination of embedded models, is developed to optimize the operation of a central air-conditioning system integrated with cold storage during fast demand response events. In [20], a smart operation strategy based on MPC is given to optimize the performance of hydronic radiant floor systems in office buildings, which includes data-driven models an optimizer based on constraint linear or quadratic programming.

(3) MPC of Microgrids

For the subject of operational optimization of microgrids, several promising MPC methods have been developed. In [21], a microgrid with solar photovoltaic (PV) and battery energy storage (BES) is considered, and a MPC-based state of charge (SOC)-oriented charging control system is developed to smooth the PV output. In [22], the optimal economic dispatch problem of combined heat and power (CHP) microgrids is solved by properly designing a MPC. In [23], a MPC based robust scheduling strategy is proposed to utilize the flexibility of Community integrated energy system (CIES) for enhancing the uncertainty adaptability.

(4) MPC of Electric Vehicles

MPC has also been applied to the field of energy management for electric vehicles or ships. To improve computational efficiency of energy management strategies for plug-in hybrid electric vehicles (PHEVs), a stochastic MPC based on PMP is proposed in [24], which differs from widely used DP-based predictive methods. In [25], a generic control methodology based on receding horizon control techniques is proposed for the ship maneuvering control as well as energy management.

(5) MPC of Nuclear Plants

To obtain a high efficiency in nuclear plant operation, the idea of MPC was also introduced to the power-level control of nuclear plants [26]. Some promising MPC algorithms such as the fuzzy MPC [27] and nonlinear MPC [28,29] were proposed for the large-range power-level

maneuvering of nuclear reactors. Further, the ANN-based [30] and dynamic matrix control (DMC)-based [31] MPC methods were also proposed for the supervisory control of nuclear steam supply systems (NSSSs), which regulate the NSSS thermal power by adjusting the set-points of local controllers that regulates the process variables such as neutron flux as well as flowrates, temperatures and pressures of primary and secondary coolants [32].

It can be seen that the prediction models are central in applying the MPC algorithms, however, the identification of the models or their parameters are usually expensive especially for those safety-critical processes such as nuclear reactors, which leads to their complexity in engineering deployment relative to those simple proportional-integral-differential (PID) controllers.

From the above introduction to the current status in the field of energy system dynamical optimization based upon both classical optimal control theory and MPC methods, it is very necessary to develop novel methods which can realize effective optimization under the existence of strong dynamical uncertainty. Actually, reinforcement learning control (RLC) method, which is inspired by natural learning mechanisms such that animals adjust their actions from the reward and punishment stimuli given by environment, enables the design of adaptive controllers through solving user-prescribed optimal control problems online with the feedback given by the real-time evaluative information from the environment [33,34]. From the viewpoint of system control, RLC bridge the gap between the classical optimal control and adaptive control methods, and has the virtues of both optimization and adaptation. Unlike the classical optimal controllers designed offline based upon the full knowledge of system dynamics, RLC approximates the solution of HJB equation online in real time without accurately knowing system models. Unlike the traditional adaptive controllers that have no optimization capability, RLC takes a goal-directed behavior that optimizes certain performance indices by action-based online learning [35]. Unlike those MPC algorithms which not only rely on either physics-based or identified prediction models but also are difficult in providing closed-loop stability, RLC is actually a model-free control designed based on the knowledge about the structure of system dynamics at most, and the stabilizing capability of RLC is inherently guaranteed by the mechanism of approximating the solution of HJB equation. The actor-critic structure shown in Fig. 1 is widely adopted in real-time implementation of RLC algorithms, wherein an actor component applies an action to the environment, and a critical component assesses the value of that action by interacting with the environment. The learning mechanism in the actor-critic structure shown in Fig. 1 is to perform policy evaluation and improvement successively, where the evaluation is executed by the critic from observing the result of applying current control action, and where the improvement is in the sense that the new control action yields a better response relative to previous action. It is shown in [36,37] that the RLC implemented through policy-iteration and value iteration schemes not only converges to the optimal control solving the HJB function but also can stabilize general nonlinear systems. Usually, the actor and the critic are realized by the linear parameterized ANNs such as the radial basis function (RBF) network without any knowledge about the system dynamics [33,34]. Nowadays, the RLC is a hot spot in the field of optimal

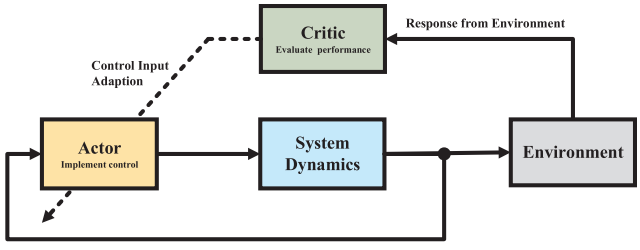

Fig. 1. The actor-critic structure of reinforcement learning control.

adaptive control, and there have been many meaningful results. In [38], a distributed ANN-based RLC was proposed through constructing a set of critic neural networks which can approximate the cost functions online and give the control policies by solving HJB equation. In [39], a reinforcement learning (RL)-based robust adaptive control algorithm was developed for a class of continuous-time uncertain nonlinear systems subject to input constraints, where the robustness is guaranteed by the appropriate selection of value functions for the nominal system. In [40], the idea of RLC is introduced into the passivity-based control (PBC) method, which does not require the specification of a global desired Hamiltonian while preserving the system's matching conditions. In [41,42], a novel RLC with actor-critic-identifier (ACI) structure using three ANNs is proposed, where the actor and critic ANNs approximate the optimal control and the optimal value function determined by HJB equation respectively, and the ANN adopted for identification can provide an asymptotically approximation to the uncertain system dynamics. The persistent excitation (PE) condition should be required by the RLC with ACI structure so that both the exponential convergence to a bounded region near the optimal control and the uniformly ultimately bounded (UUB) closed-loop stability can be guaranteed. However, it is generally impossible to guarantee the PE condition a priori, and hence, a probing signal using trial and error should be added to the controller to ensure PE, which might induce the deterioration of optimality and stability. To avoid adding probing signals while keeping UUB closed-loop stability and the convergence to a neighborhood of the optimal policy, model-based RLCs were recently developed [43,44], which can be implemented by evaluating the Bellman error at a number of points in the state space of a system model tuned by a concurrent learning (CL)-based identifier.

Besides the theoretical results on RLC, RLC designs have been applied for practical online control optimization. In [45], an improved adaptive RLC is designed for unmanned air vehicle morphing control, where the RLC learning the optimal shape change policy is integrated with an adaptive dynamical inversion tracking control. In [46], a RLC approach is proposed for PV/T array and the full building model, and it was shown by simulation results that the RLC outperforms a rule-based control (RBC) by over 10% after the third year of implementation. In [47], a deep reinforcement learning (DRL) based data-driven model-free load-frequency control (LFC) against renewable generation uncertainties is given for minimizing frequency deviation with faster response and stronger adaptability for unmolded dynamics. In [48], the RLC method is applied to realize adaptive traffic signal control (ATSC) that has strong potential to effectively alleviate urban traffic congestion, and a novel multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC) were proposed. In [49], a novel RLC called interleaved learning control is given for the real-time solution of the optimal operational control problem of the industrial flotation process which a key component in the mineral processing concentrator line. In [50], the RLC method is applied to an intelligent human-robot interaction (HRI) system with adjustable robot behavior, and the problem of optimal parameter tuning of the prescribed robot impedance is transformed to a linear quadratic regulation (LQR) problem which minimizes the human effort and optimizes closed-loop behavior. To obviate the requirement of accurate human model, an integral reinforcement learning mechanism was used to solve the LQR problem. In [51], a deep reinforcement learning-based control-aware scheduling (DeepCAS) algorithm is developed to tackle the scheduling issues induced by the fact that the size of communication network is smaller than the size of complex systems such as the internet of things (IoT) and large-scale industrial processes.

From the above introduction to the mechanism of RLC as well as the comparisons to the methods of classical optimal control, adaptive control and MPC, it can be seen that RLC is suitable for the dynamical optimization of complex energy systems. Actually, the dynamic optimization methods of energy systems are usually implemented as the supervisory controllers. Supervisory controllers have a long tradition in

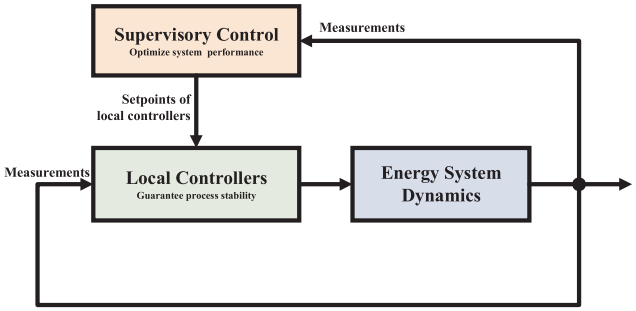

Fig. 2. Relationship between local controllers and supervisory control.

energy system control engineering practice, which deal with distributing global energy demands of the plant to individual demands required by each member of a set of service equipment, such as reactors, boilers, heat-exchangers, pumps, blowers, valves and the like, while minimizing the predetermined costs. Usually, the individual demands are translated into the set-points of local PID controllers which are well tuned for closed-loop stability. The relationship between the supervisory controller and local controllers are shown in Fig. 2. Due to the closed-loop stability of complex industrial energy processes such as the fossil, chemical and nuclear plants are guaranteed by local PID controllers, the supervisory controller actually handles the optimal control problem of nonlinear dissipative systems which can be solved by some properly developed RLC methods.

In this paper, motivated by the necessity of developing dynamical optimization method for complex energy systems, a multi-layer perception (MLP)-based RLC is proposed for the nonlinear dissipative systems governing the dynamics coupled by energy systems and their local controllers, which is able to optimize a given performance index online without knowing the accurate system dynamics. The main contributions are given as follows:

(1) An MLP-based state-observer is given for dynamical identification, which converges to a bounded neighborhood of the actual dynamics asymptotically, and takes the form as a linear system with its state and input matrices as the functions of system state-vector and MLP weighting matrices.
(2) An approximated optimal control is given by solving the algebraic Riccati equation related to the MLP-based state-observer, which effectively approximates the optimal control corresponding to a given performance index online.
(3) It is proven that the MLP-based RLC interconnected by both the MLP-based state-observer and the approximated optimal controller can guarantee uniformly ultimately bounded (UUB) closed-loop stability.
(4) This newly-built MLP-based RLC is applied to optimize the thermal power response of a nuclear steam supply system (NSSS) based upon the high temperature gas-cooled reactor (HTGR). Simulation results show not only the satisfactory performance but also the influences from the controller parameters to closed-loop responses.

It is worthy to be noted here that the MLP-based RLC proposed in this paper is NOT utilized for local control, i.e. to regulate the process variables such as pressure, temperature and flowrate. Actually, this MLP-based RLC can be adopted as a supervisory control to optimize energy system operation by properly generating the setpoints of local controllers, where the inputs are the measurements of the process variables needed to be optimized and the outputs are the local setpoints. It can be seen that if the MLP-based RLC is used to optimize the response of a single process variable, then it is a single-input-single-output (SISO) supervisory controller. If this RLC is used to optimize the responses of multiple process variables, then it is a multi-inputmultioutput (MIMO) supervisory controller.

2. Problem formulation

Since supervisory controller of a given energy system deals with the dynamic system interconnected by energy generation or consumption processes and their local controllers, consider the nonlinear dissipative system taking the form as

$$\dot{\boldsymbol{x}} = -\boldsymbol{\Omega x} + \boldsymbol{f}_0(\boldsymbol{x}) + \sum_{i=1}^{m} \boldsymbol{f}_i(\boldsymbol{x}) u_i \tag{1}$$

where $\boldsymbol{x} \in \mathrm{R}^n$ and $\boldsymbol{u} = [u_1, \ldots, u_m]^{\mathrm{T}} \in \mathrm{R}^m$ are the state-vector and control input respectively, $\boldsymbol{f}_k$ ($k = 0,1, \ldots, m$) is a norm-bounded continuously differentiable vector-valued function, $\Omega = \mathrm{diag}([\omega_1, \ldots, \omega_n])$ is a given positive-definite diagonal matrix. It is assumed that system (1) is controllable, state-vector $\boldsymbol{x}$ can be directly measured, and control input $\boldsymbol{u}$ is norm-bounded.

Define a smooth cost function associated with feedback control law

$$\boldsymbol{u} = \boldsymbol{\mu}(\boldsymbol{x}) \tag{2}$$

as

$$J(\boldsymbol{x}) = \int_t^{\infty} r(\boldsymbol{x}(\tau), \boldsymbol{u}(\tau)) \mathrm{d}\tau, \tag{3}$$

where $\boldsymbol{\mu}$ is a continuous vector-valued function, r is the utility given by

$$r(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q x} + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{R u}, \tag{4}$$

$\boldsymbol{Q} = \boldsymbol{Q}^{\mathrm{T}}$ and $\boldsymbol{R} = \boldsymbol{R}^{\mathrm{T}}$ are a positive-definite matrices. From Leibniz's formula, the infinitesimal equivalence of (3) can be written as

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q x} + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{R u} + [\nabla J(\boldsymbol{x})]^{\mathrm{T}} [\boldsymbol{f}_0(\boldsymbol{x}) + \boldsymbol{B}(\boldsymbol{x}) \boldsymbol{u} - \boldsymbol{\Omega x}] = 0, \tag{5}$$

which is called the continuous-time Bellman equation. Here, ∇J is a column vector denoting the gradient of cost function J, and

$$\boldsymbol{B}(\boldsymbol{x}) = [\boldsymbol{f}_1(\boldsymbol{x}) \ \cdots \ \boldsymbol{f}_m(\boldsymbol{x})]. \tag{6}$$

Further, define Hamiltonian function H as

$$H(\boldsymbol{x}, \boldsymbol{u}, \nabla J) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q x} + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{R u} + [\nabla J(\boldsymbol{x})]^{\mathrm{T}} [\boldsymbol{f}_0(\boldsymbol{x}) + \boldsymbol{B}(\boldsymbol{x}) \boldsymbol{u} - \boldsymbol{\Omega x}]. \tag{7}$$

It can be seen that the optimal control satisfies

$$\boldsymbol{\mu}^* = \arg\min_{\mu} H(\boldsymbol{x}, \boldsymbol{\mu}(\boldsymbol{x}), \nabla J), \tag{8}$$

and from equation (7),

$$\boldsymbol{\mu}^* = -\frac{1}{2} \boldsymbol{R}^{-1} \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}) \nabla J(\boldsymbol{x}). \tag{9}$$

Substitute equation (9) back to (5), it can be seen that $J(\boldsymbol{x})$ is the solution of

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q x} - \frac{1}{4} \nabla J^{\mathrm{T}}(\boldsymbol{x}) \boldsymbol{B}(\boldsymbol{x}) \boldsymbol{R}^{-1} \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}) \nabla J(\boldsymbol{x}) + [\nabla J(\boldsymbol{x})]^{\mathrm{T}} [\boldsymbol{f}_0(\boldsymbol{x}) - \boldsymbol{\Omega x}] = 0, \tag{10}$$

which is called Hamilton-Jacobi-Bellman (HJB) equation.

Furthermore, suppose there exists a positive-definite matrix $\boldsymbol{P}$ so that

$$J(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P x}. \tag{11}$$

By substituting (11) to (10) and (9), the HJB equation and optimal controller can be rewritten as

$$\boldsymbol{x}^{\mathrm{T}} [\boldsymbol{Q} - \boldsymbol{P B}(\boldsymbol{x}) \boldsymbol{R}^{-1} \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}) \boldsymbol{P}] \boldsymbol{x} + 2 \boldsymbol{x}^{\mathrm{T}} \boldsymbol{P} [\boldsymbol{f}_0(\boldsymbol{x}) - \boldsymbol{\Omega x}] = 0. \tag{12}$$

and

$$\boldsymbol{\mu}^* = -\boldsymbol{R}^{-1} \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{x}) \boldsymbol{P x} \tag{13}$$

respectively. If HJB equation (12) can be solved online, then we can apply optimal control (13) so as to minimize performance index (3).

3. MLP-based approximation of optimal control

3.1. MLP-based approximation of system dynamics

Since the MLP with one or more hidden layers is able to approximate any continuous nonlinear function [36], there exists $m + 1$ MLPs with two hidden layers

$$\boldsymbol{f}_{k,\mathrm{MLP}}(\boldsymbol{W}_k, \boldsymbol{V}_k, \boldsymbol{x}) = \boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{x}), \quad k = 0, 1, \cdots, m, \tag{14}$$

so that norm-bounded vector valued function $\boldsymbol{f}_k$ ($k = 0,1, \ldots, m$) in model (1) can be represented by

$$\boldsymbol{f}_k(\boldsymbol{x}) = \boldsymbol{f}_{k,\mathrm{MLP}}(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{x}) + \boldsymbol{d}_k = \boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{x}) + \boldsymbol{d}_k, \tag{15}$$

where $k = 0, 1, \ldots, m$, norm-bounded vector $\boldsymbol{d}_k \in \mathrm{R}^n$ is the approximation error satisfying

$$\boldsymbol{d}_k = \boldsymbol{f}_k(\boldsymbol{x}) - \boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{x}), \tag{16}$$

$\boldsymbol{V}_k \in \mathrm{R}^{n\times l}$ and $\boldsymbol{W}_k \in \mathrm{R}^{l\times n}$ are respectively the first-to-second and second-to-third layer interconnection weight matrices,

$$\boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{x}) = [s(\boldsymbol{v}_{k,1}^{\mathrm{T}}\boldsymbol{x}) \;\; s(\boldsymbol{v}_{k,2}^{\mathrm{T}}\boldsymbol{x}) \;\; \cdots \;\; s(\boldsymbol{v}_{k,l}^{\mathrm{T}}\boldsymbol{x})]^{\mathrm{T}}, \tag{17}$$

$\boldsymbol{v}_{k,i} \in \mathrm{R}^l$ ($i = 1, \ldots, n$) is the ith column of matrix $\boldsymbol{V}_k$, and s is the sigmoid function given by

$$s(z) = \frac{1}{1 + e^{-z}}, \; z \in \mathrm{R}. \tag{18}$$

The approximation capability of MLP (14) is shown by the following lemma.

Lemma 1 (*[52]*). *Consider three layer MLPs given by equation (14). For an arbitrarily given $D_k > 0$ ($k = 0, 1, \ldots, m$), there exits l, $\boldsymbol{V}_k$ and $\boldsymbol{W}_k$ such that approximation error $\boldsymbol{d}_k$ defined in (16) satisfies*

$\|\boldsymbol{d}_k\|_2 < D_k$.

Remark 1. Based on Lemma 1, it can be seen that the MLPs can give effective approximations to the vector functions $\boldsymbol{f}_k$ ($k = 0, 1, \ldots, m$) in dynamical equation (1), and the approximation error can be arbitrarily small. Then, system dynamics (1) can be represented by the MLPs as

$$\dot{\boldsymbol{x}} = -\boldsymbol{\Omega}\boldsymbol{x} + \boldsymbol{W}_0^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x}) + \sum_{k=1}^{m}\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{x})u_i + \boldsymbol{d}, \tag{19}$$

where $\boldsymbol{d} = \boldsymbol{d}_0 + \boldsymbol{d}_1 + \cdots + \boldsymbol{d}_m$ is also norm-bounded.

Remark 2. From Lemma 1, since there exists l, $\boldsymbol{V}_k$ and $\boldsymbol{W}_k$ so that $||\boldsymbol{d}||_2$ is arbitrarily small, the MLP-approximation of system dynamics (1) can be written as

$$\dot{\boldsymbol{x}} = -\boldsymbol{\Omega}\boldsymbol{x} + \sum_{k=0}^{m}\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{x})u_i, \tag{20}$$

with u_0 defined as $u_0 = 1$. Further, suppose that the performance index of MLP-approximated dynamics still satisfies equations (3) and (12). Then the corresponding HJB equation as well as the approximated optimal control are given by

$$\boldsymbol{x}^{\mathrm{T}}[\boldsymbol{Q} - \boldsymbol{P}\boldsymbol{B}_{\mathrm{NN}}(\boldsymbol{x})\boldsymbol{R}^{-1}\boldsymbol{B}_{\mathrm{NN}}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{P}]\boldsymbol{x} + 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}[\boldsymbol{W}_0^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x}) - \boldsymbol{\Omega}\boldsymbol{x}] = 0 \tag{21}$$

and

$$\boldsymbol{\mu}_{\mathrm{A}}^{*} = -\boldsymbol{R}^{-1}\boldsymbol{B}_{\mathrm{NN}}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{P}\boldsymbol{x}, \tag{22}$$

respectively, where positive-definite symmetric matrix $\boldsymbol{P}$ is the solution of HJB equation (21), and matrix-valued function $\boldsymbol{B}_{\mathrm{NN}}(\boldsymbol{x})$ satisfying equation

$$\boldsymbol{B}_{\mathrm{NN}}(\boldsymbol{x}) = [\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_1^{\mathrm{T}}\boldsymbol{x}) \;\; \boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_2^{\mathrm{T}}\boldsymbol{x}) \;\; \cdots \;\; \boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_m^{\mathrm{T}}\boldsymbol{x})]. \tag{23}$$

is the MLP-based approximation of system input matrix $\boldsymbol{B}(\boldsymbol{x})$ defined by (6). The key to design approximated optimal control (22) is to give the solution $\boldsymbol{P}$ of HJB equation (21), which is the central in following Section 3.2.

3.2. Design of approximated optimal control

In this subsection, it is shown that HJB equation (21) is equivalent to an algebraic Riccati equation, which is much easy to be solved to obtain matrix $\boldsymbol{P}$ in approximated optimal control (22). This result is summarized as following Proposition 1 which is the first main result of this paper.

Proposition 1. *Consider nonlinear dissipative system (1) with performance index $J(\boldsymbol{x})$ defined by (3) and utility function $r(\boldsymbol{x}, \boldsymbol{u})$ given by (4). Suppose that performance index J can be represented as (11), and then HJB equation (21) is equivalent to algebraic Riccati equation*

$$\boldsymbol{P}[\boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x}) - \boldsymbol{\Omega}] + [\boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x}) - \boldsymbol{\Omega}]^{\mathrm{T}}\boldsymbol{P} + \boldsymbol{Q} - \boldsymbol{P}\boldsymbol{B}_{\mathrm{NN}}(\boldsymbol{x})\boldsymbol{R}^{-1}\boldsymbol{B}_{\mathrm{NN}}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{P} = 0, \tag{24}$$

where

$$\boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x}) = \boldsymbol{W}_0^{\mathrm{T}}\boldsymbol{U}_0^{-1}(\boldsymbol{V}_0, \boldsymbol{x})\boldsymbol{S}'(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x})\boldsymbol{V}_0^{\mathrm{T}}, \tag{25}$$

$$\boldsymbol{U}_0(\boldsymbol{V}_0, \boldsymbol{x}) = \mathrm{diag}\left(\begin{bmatrix} \boldsymbol{v}_{0,1}^{\mathrm{T}}\boldsymbol{x}(1 - s(\boldsymbol{v}_{0,1}^{\mathrm{T}}\boldsymbol{x})) & \boldsymbol{v}_{0,2}^{\mathrm{T}}\boldsymbol{x}(1 - s(\boldsymbol{v}_{0,2}^{\mathrm{T}}\boldsymbol{x})) & \cdots & \boldsymbol{v}_{0,l}^{\mathrm{T}}\boldsymbol{x}(1 - s(\boldsymbol{v}_{0,l}^{\mathrm{T}}\boldsymbol{x})) \end{bmatrix}\right), \tag{26}$$

and $\boldsymbol{B}_{\mathrm{NN}}$(x) is defined by (23).

Proof. From HJB equation (21), if MLP $\boldsymbol{W}_0^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x})$, which is utilized to approximate vector-valued function $\boldsymbol{f}_0$ in system (1), can be rewritten as

$$\boldsymbol{W}_0^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x}) = \boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x})\boldsymbol{x}, \tag{27}$$

where $\boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x})$ is a matrix-valued function of state-vector $\boldsymbol{x}$ and MLP weighting matrices $\boldsymbol{W}_0$ and $\boldsymbol{V}_0$, then HJB equation (21) can be transformed to algebraic Riccati equation (24). The following parts of this proof focuses on finding a proper $\boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x})$.

From equation (18), it can be derived from direct differentiation that

$$s'(z) = \frac{\mathrm{d}s}{\mathrm{d}z} = \frac{e^{-z}}{(1 + e^{-z})^2} = [1 - s(z)]s(z). \tag{28}$$

By defining

$$\boldsymbol{S}'(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x}) = [s'(\boldsymbol{v}_{0,1}^{\mathrm{T}}\boldsymbol{x}) \;\; s'(\boldsymbol{v}_{0,2}^{\mathrm{T}}\boldsymbol{x}) \;\; \cdots \;\; s'(\boldsymbol{v}_{0,l}^{\mathrm{T}}\boldsymbol{x})]^{\mathrm{T}}, \tag{29}$$

we have

$$\begin{aligned}\boldsymbol{S}'(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x})\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x} &= [\boldsymbol{v}_{0,1}^{\mathrm{T}}\boldsymbol{x}s'(\boldsymbol{v}_{0,1}^{\mathrm{T}}\boldsymbol{x}) \;\; \boldsymbol{v}_{0,2}^{\mathrm{T}}\boldsymbol{x}s'(\boldsymbol{v}_{0,2}^{\mathrm{T}}\boldsymbol{x}) \;\; \cdots \;\; \boldsymbol{v}_{0,l}^{\mathrm{T}}\boldsymbol{x}s'(\boldsymbol{v}_{0,l}^{\mathrm{T}}\boldsymbol{x})]^{\mathrm{T}} \\ &= \boldsymbol{U}_0(\boldsymbol{V}_0, \boldsymbol{x})\boldsymbol{S}(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x}),\end{aligned} \tag{30}$$

where $\boldsymbol{U}_0$ is defined by equation (26). From equation (30), if matrix $\boldsymbol{U}_0$ is nonsingular, then we can express $\boldsymbol{S}$ as

$$\boldsymbol{S}(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x}) = \boldsymbol{U}_0^{-1}(\boldsymbol{V}_0, \boldsymbol{x})\boldsymbol{S}'(\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x})\boldsymbol{V}_0^{\mathrm{T}}\boldsymbol{x}. \tag{31}$$

Through substituting equations (31) to (27), it can be seen that $\boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x})$ can be chosen as the form given by equation (25), which means that HJB equation (21) is equivalent to algebraic Riccati equation (24). Further, matrix $\boldsymbol{P}$ utilized to determine approximated optimal control (22) can be obtained by solving algebraic Riccati equation (24) instead of HJB equation (21). This completes the proof of Proposition 1.

Remark 3. Based on equation (27), MLP-approximated dynamics (20) can be rewritten as

$$\dot{\boldsymbol{x}} = [\boldsymbol{A}_{\mathrm{NN}}(\boldsymbol{x}) - \boldsymbol{\Omega}]\boldsymbol{x} + \boldsymbol{B}_{\mathrm{NN}}(\boldsymbol{x})\boldsymbol{u}, \tag{32}$$

where $\boldsymbol{B}_{\mathrm{NN}}$ is given by (23). Since system (1) is assumed to be controllable, MLP-approximated system (32) is also controllable.

Moreover, with comparison to the current RLC methods, the feedback gain is given by solving an algebraic Riccati equation given by the state-vector and the MLP-approximated model, which is much simpler and more implementable.

4. Adaptation mechanism of MLP weighting matrices

From (25) and (23), solution $\boldsymbol{P}$ of algebraic Riccati equation (24) is determined by state-vector $\boldsymbol{x}$ and weighting matrices $\boldsymbol{W}_k$ and $\boldsymbol{V}_k$ of MLPs. However, since both $\boldsymbol{W}_k$ and $\boldsymbol{V}_k$ are not known in prior, the approximated optimal control given by (22) and (24) cannot be practically implemented. This leads to the necessity of developing the learning laws that can give the estimations of $\boldsymbol{W}_k$ and $\boldsymbol{V}_k$, which is given in this section.

Design an estimator of system (19) as

$$\dot{\hat{\boldsymbol{x}}} = -\boldsymbol{\Omega}\hat{\boldsymbol{x}} + \sum_{k=0}^{m} \widehat{\boldsymbol{W}}_k^{\mathrm{T}} \boldsymbol{S}(\widehat{\boldsymbol{V}}_k^{\mathrm{T}}\hat{\boldsymbol{x}}) u_k - \boldsymbol{H}\boldsymbol{e}, \tag{33}$$

where

$$\boldsymbol{e} = \hat{\boldsymbol{x}} - \boldsymbol{x}, \tag{34}$$

$\hat{\boldsymbol{x}} \in \mathrm{R}^n$ is the estimation of $\boldsymbol{x}$, $\widehat{\boldsymbol{V}}_k \in \mathrm{R}^{n\times l}$ and $\widehat{\boldsymbol{W}}_k \in \mathrm{R}^{l\times n}$ are respectively the estimations of weight matrices $\boldsymbol{V}_k$ and $\boldsymbol{W}_k$, $\boldsymbol{H} \in \mathrm{R}^{n\times n}$ is the observer gain matrix, and $k = 0, 1, \cdots, m$. In the following of this subsection, the adaptation laws for updating $\widehat{\boldsymbol{V}}_k$ and $\widehat{\boldsymbol{W}}_k$ online are proposed so that estimation error $\boldsymbol{e}$ is uniformly ultimately bounded (UUB).

First, a useful lemma is introduced to obtain the dynamics of estimation error $\boldsymbol{e}$.

Lemma 2 (*[52]*). *Define*

$$\boldsymbol{G}_{\mathrm{E},k}(\widehat{\boldsymbol{W}}_k, \widehat{\boldsymbol{V}}_k) = \widehat{\boldsymbol{W}}_k^{\mathrm{T}} \boldsymbol{S}(\widehat{\boldsymbol{V}}_k^{\mathrm{T}}\hat{\boldsymbol{x}}) - \boldsymbol{W}_k^{\mathrm{T}} \boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\hat{\boldsymbol{x}}), \tag{35}$$

for $k = 0,1, \ldots, m$. It can be seen that function $\boldsymbol{G}_E$ satisfies

$$\boldsymbol{G}_{\mathrm{E},k}(\widehat{\boldsymbol{W}}_k, \widehat{\boldsymbol{V}}_k) = \widetilde{\boldsymbol{W}}_k^{\mathrm{T}}(\widehat{\boldsymbol{S}}_k - \widehat{\boldsymbol{S}}'_k \widehat{\boldsymbol{V}}_k^{\mathrm{T}}\hat{\boldsymbol{x}}) + \widehat{\boldsymbol{W}}^{\mathrm{T}} \widehat{\boldsymbol{S}}'_k \widetilde{\boldsymbol{V}}_k^{\mathrm{T}}\hat{\boldsymbol{x}} + \boldsymbol{d}_{\mathrm{E},k} \tag{36}$$

where

$$\widetilde{\boldsymbol{W}}_k = \widehat{\boldsymbol{W}}_k - \boldsymbol{W}_k, \tag{37}$$

$$\widetilde{\boldsymbol{V}}_k = \widehat{\boldsymbol{V}}_k - \boldsymbol{V}_k, \tag{38}$$

$$\widehat{\boldsymbol{S}}_k = \boldsymbol{S}(\widehat{\boldsymbol{V}}_k^{\mathrm{T}}\hat{\boldsymbol{x}}), \tag{39}$$

$$\widehat{\boldsymbol{S}}'_k = \boldsymbol{S}'(\widehat{\boldsymbol{V}}_k^{\mathrm{T}}\hat{\boldsymbol{x}}), \tag{40}$$

and $\boldsymbol{d}_{E,k}$ is the norm-bounded approximation error.

Based upon Lemma 2 and by subtracting MLP-approximated dynamics (20) from observer dynamics (33), the dynamics of observation error $\boldsymbol{e}$ can be given by

$$\dot{\boldsymbol{e}} = -\boldsymbol{\Omega}\boldsymbol{e} + \sum_{k=0}^{m} u_k \boldsymbol{G}_{\mathrm{E},k}(\widehat{\boldsymbol{W}}_k, \widehat{\boldsymbol{V}}_k) - \overline{\boldsymbol{d}}, \tag{41}$$

where function $\boldsymbol{G}_{\mathrm{E},k}$ is defined by (35), and

$$\overline{\boldsymbol{d}} = \boldsymbol{d} + \sum_{k=0}^{m} u_k \boldsymbol{W}_k^{\mathrm{T}}[\boldsymbol{S}(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{x}) - \widehat{\boldsymbol{S}}_k]. \tag{42}$$

From Lemma 1 and the boundness of control input u, $\overline{\boldsymbol{d}}$ is also norm-bounded.

It can be seen that if observation error dynamics (41) is stable, then state-observer (33) gives an effective reconstruction of state-vector $\boldsymbol{x}$. It is shown in the following Proposition 2 that properly designed learning laws of weighting matrices $\widehat{\boldsymbol{V}}_k$ and $\widehat{\boldsymbol{W}}_k$ ($k = 0, 1, \cdots, m$) can guarantee a globally bounded estimation error $\boldsymbol{e}$, which is the second main result of this paper.

Proposition 2. *The MLP-based prediction model (33) with weighting matrices $\widehat{\boldsymbol{V}}_k$ and $\widehat{\boldsymbol{W}}_k$ updated online by learning laws*

$$\dot{\widehat{\boldsymbol{W}}}_k = -u_k \Gamma_{\mathrm{W},k}[\widehat{\boldsymbol{S}}_k - \widehat{\boldsymbol{S}}'_k \widehat{\boldsymbol{V}}_k^{\mathrm{T}}\hat{\boldsymbol{x}}]\boldsymbol{e}^{\mathrm{T}} \tag{43}$$

and

$$\dot{\widehat{\boldsymbol{V}}}_k = -u_k \Gamma_{\mathrm{V},k}\boldsymbol{x}\boldsymbol{e}^{\mathrm{T}}\widehat{\boldsymbol{W}}_k \widehat{\boldsymbol{S}}'_k \tag{44}$$

can provide a UUB estimation error $\boldsymbol{e}$, where both $\Gamma_{W,k}$ and $\Gamma_{V,k}$ are given diagonal positive-definite matrices, and $k = 0,1, \ldots, m$.

Proof. To analyze the UUB stability of estimation error $\boldsymbol{e}$, choose the Lyapunov function as

$$V_{\mathrm{E}}(\boldsymbol{e}, \widetilde{\boldsymbol{W}}_k, \widetilde{\boldsymbol{V}}_k) = \frac{1}{2}\left\{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} + \sum_{k=0}^{m}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_k^{\mathrm{T}}\Gamma_{\mathrm{W},k}^{-1}\widetilde{\boldsymbol{W}}_k) + \mathrm{tr}(\widetilde{\boldsymbol{V}}_k^{\mathrm{T}}\Gamma_{\mathrm{V},k}^{-1}\widetilde{\boldsymbol{V}}_k)]\right\}, \tag{45}$$

where function tr(·) is the trace of a matrix.

Differentiate of function V_{E} along the trajectory of observation error dynamics can be given by (41)

$$\begin{aligned}\dot{V}_{\mathrm{E}}(\boldsymbol{e}, \widetilde{\boldsymbol{W}}_k, \widetilde{\boldsymbol{V}}_k) &= \boldsymbol{e}^{\mathrm{T}}\dot{\boldsymbol{e}} + \textstyle\sum_{k=0}^{m}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_k^{\mathrm{T}}\Gamma_{\mathrm{W},k}^{-1}\dot{\widehat{\boldsymbol{W}}}_k) + \mathrm{tr}(\widetilde{\boldsymbol{V}}_k^{\mathrm{T}}\Gamma_{\mathrm{V},k}^{-1}\dot{\widehat{\boldsymbol{V}}}_k)] \\ &= -\boldsymbol{e}^{\mathrm{T}}(\boldsymbol{\Omega} + \boldsymbol{H})\boldsymbol{e} + \boldsymbol{e}^{\mathrm{T}}[\textstyle\sum_{k=0}^{m} u_k \boldsymbol{G}_{\mathrm{E},k}(\widehat{\boldsymbol{W}}_k, \widehat{\boldsymbol{V}}_k) - \overline{\boldsymbol{d}}] \\ &\quad + \textstyle\sum_{k=0}^{m}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_k^{\mathrm{T}}\Gamma_{\mathrm{W},k}^{-1}\dot{\widehat{\boldsymbol{W}}}_k) + \mathrm{tr}(\widetilde{\boldsymbol{V}}_k^{\mathrm{T}}\Gamma_{\mathrm{V},k}^{-1}\dot{\widehat{\boldsymbol{V}}}_k)] \\ &= -\boldsymbol{e}^{\mathrm{T}}(\boldsymbol{\Omega} + \boldsymbol{H})\boldsymbol{e} + \textstyle\sum_{k=0}^{m}\mathrm{tr}\{\widetilde{\boldsymbol{W}}_k^{\mathrm{T}}\Gamma_{\mathrm{W},k}^{-1}[\dot{\widehat{\boldsymbol{W}}}_k + u_k\Gamma_{\mathrm{W},k}(\widehat{\boldsymbol{S}}_k - \widehat{\boldsymbol{S}}'_k\widehat{\boldsymbol{V}}_k^{\mathrm{T}}\boldsymbol{x})\boldsymbol{e}^{\mathrm{T}}]\} \\ &\quad + \textstyle\sum_{k=0}^{m}\mathrm{tr}\{\widetilde{\boldsymbol{V}}_k^{\mathrm{T}}\Gamma_{\mathrm{V},k}^{-1}(\dot{\widehat{\boldsymbol{V}}}_k + u_k\Gamma_{\mathrm{V}}\boldsymbol{Z}\boldsymbol{e}^{\mathrm{T}}\widehat{\boldsymbol{W}}\widehat{\boldsymbol{S}}')\} + \boldsymbol{e}^{\mathrm{T}}\boldsymbol{d}_{\mathrm{A}}\end{aligned} \tag{46}$$

where

$$\boldsymbol{d}_{\mathrm{A}} = \boldsymbol{d}_{\mathrm{E}} - \overline{\boldsymbol{d}}, \tag{47}$$

$$\boldsymbol{d}_{\mathrm{E}} = \sum_{k=0}^{m} u_k \boldsymbol{d}_{\mathrm{E},k}, \tag{48}$$

and it can be seen that $\boldsymbol{d}_{\mathrm{A}}$ is a norm-bounded vector denoting the approximation error.

Then, substitute adaptation laws (43) and (44) to equation (46),

$$\begin{aligned}\dot{V}_{\mathrm{E}}(\boldsymbol{e}, \widetilde{\boldsymbol{W}}_k, \widetilde{\boldsymbol{V}}_k) &= -\boldsymbol{e}^{\mathrm{T}}(\boldsymbol{\Omega} + \boldsymbol{H})\boldsymbol{e} + \boldsymbol{e}^{\mathrm{T}}\boldsymbol{d}_{\mathrm{A}} \leqslant -\boldsymbol{e}^{\mathrm{T}}\left(\boldsymbol{\Omega} + \frac{1}{2}\boldsymbol{H}\right)\boldsymbol{e} + \frac{1}{2}\boldsymbol{d}_{\mathrm{A}}^{\mathrm{T}} \\ &\quad \boldsymbol{H}^{-1}\boldsymbol{d}_{\mathrm{A}} \leqslant -\frac{1}{2}\left[\lambda_{\min}(2\boldsymbol{\Omega} + \boldsymbol{H})\|\boldsymbol{e}\|_2^2 - \frac{\|\boldsymbol{d}_{\mathrm{A}}\|_2^2}{\lambda_{\min}(\boldsymbol{H})}\right].\end{aligned} \tag{49}$$

From inequality (49), estimation error $\boldsymbol{e}$ finally enters to the bounded set defined by

$$\Xi_{\mathrm{E}} = \left\{\boldsymbol{e}\ \middle|\ \|\boldsymbol{e}\|_2 \leqslant \frac{\|\boldsymbol{d}_{\mathrm{A}}\|_2}{\sqrt{\lambda_{\min}(2\boldsymbol{\Omega} + \boldsymbol{H})\lambda_{\min}(\boldsymbol{H})}}\right\}. \tag{50}$$

which means that MLP-based observer (33) provides a UUB estimation for system state $\boldsymbol{x}$. From (50), it can be further seen that the estimation error can be controlled by observer gain matrix $\boldsymbol{H}$, if the gain is higher, i.e. $\lambda_{\min}(\boldsymbol{H})$ is larger, the estimation error is smaller. This completes the proof of this proposition.

5. Reinforcement learning control design with stability analysis

In this section, based on prediction model (33) given in Section 2, a MLP-based MPC is proposed based on optimizing a given performance index subjected to the constraint of control input saturation. Moreover, the closed-loop stability of applying this newly-built MPC is analyzed.

5.1. MLP-based reinforcement learning control

Based on Propositions 1 and 2, the MLP-based reinforcement learning control (RLC) for nonlinear dissipative system (1) can be summarized as following Algorithm 1.

Algorithm 1. The reinforcement learning control of nonlinear dissipative system (1) based upon the MLP approximation of system dynamics given by observer (33) can be given as

$$\boldsymbol{u} = -\boldsymbol{R}^{-1}\widehat{\boldsymbol{B}}_{\mathrm{NN}}^{\mathrm{T}}(\widehat{\boldsymbol{x}})\widehat{\boldsymbol{P}}\widehat{\boldsymbol{x}} \tag{51}$$

where

$$\widehat{\boldsymbol{P}}\,[\widehat{\boldsymbol{A}}_{\mathrm{NN}}(\widehat{\boldsymbol{x}}) - \boldsymbol{\Omega}] + [\widehat{\boldsymbol{A}}_{\mathrm{NN}}(\widehat{\boldsymbol{x}}) - \boldsymbol{\Omega}]^{\mathrm{T}}\widehat{\boldsymbol{P}} + \boldsymbol{Q} - \widehat{\boldsymbol{P}}\widehat{\boldsymbol{B}}_{\mathrm{NN}}(\widehat{\boldsymbol{x}})\boldsymbol{R}^{-1}\boldsymbol{B}_{\mathrm{NN}}^{\mathrm{T}}(\widehat{\boldsymbol{x}})\widehat{\boldsymbol{P}} = 0 \tag{52}$$

$$\widehat{\boldsymbol{A}}_{\mathrm{NN}}(\widehat{\boldsymbol{x}}) = \widehat{\boldsymbol{W}}_0^{\mathrm{T}}\widehat{\boldsymbol{U}}_0^{-1}(\widehat{\boldsymbol{V}}_0, \widehat{\boldsymbol{x}})\widehat{\boldsymbol{S}}'_0\widehat{\boldsymbol{V}}_0^{\mathrm{T}}, \tag{53}$$

$$\widehat{\boldsymbol{B}}_{\mathrm{NN}}(\boldsymbol{x}) = \left[\widehat{\boldsymbol{W}}_1^{\mathrm{T}}\widehat{\boldsymbol{S}}_1 \;\; \boldsymbol{W}_2^{\mathrm{T}}\widehat{\boldsymbol{S}}_2 \;\; \cdots \;\; \boldsymbol{W}_m^{\mathrm{T}}\widehat{\boldsymbol{S}}_m\right], \tag{54}$$

$$\widehat{\boldsymbol{U}}_0(\widehat{\boldsymbol{V}}_0, \widehat{\boldsymbol{x}}) = \mathrm{diag}\left(\begin{bmatrix} \widehat{\boldsymbol{v}}_{0,1}^{\mathrm{T}}\widehat{\boldsymbol{x}}\,(1 - s(\widehat{\boldsymbol{v}}_{0,1}^{\mathrm{T}}\widehat{\boldsymbol{x}})) & \widehat{\boldsymbol{v}}_{0,2}^{\mathrm{T}}\widehat{\boldsymbol{x}}\,(1 - s(\widehat{\boldsymbol{v}}_{0,2}^{\mathrm{T}}\widehat{\boldsymbol{x}})) & \cdots & \widehat{\boldsymbol{v}}_{0,l}^{\mathrm{T}}\widehat{\boldsymbol{x}}\,(1 - s(\widehat{\boldsymbol{v}}_{0,l}^{\mathrm{T}}\widehat{\boldsymbol{x}})) \end{bmatrix}\right) \tag{55}$$

and state-estimation $\widehat{\boldsymbol{x}}$ as well as weighing matrices $\widehat{\boldsymbol{W}}_k$ and $\widehat{\boldsymbol{V}}_k$ are updated by MLP-based observer (33) and learning laws (43) and (44), i.e.

$$\begin{cases} \dot{\widehat{\boldsymbol{x}}} = -\boldsymbol{\Omega}\widehat{\boldsymbol{x}} + \sum_{k=0}^{m}\widehat{\boldsymbol{W}}_k^{\mathrm{T}}\boldsymbol{S}(\widehat{\boldsymbol{V}}_k^{\mathrm{T}}\widehat{\boldsymbol{x}})u_k - \boldsymbol{H}\boldsymbol{e}, \\ \dot{\widehat{\boldsymbol{W}}}_k = -u_k\Gamma_{\mathrm{W},k}[\widehat{\boldsymbol{S}}_k - \widehat{\boldsymbol{S}}'_k\widehat{\boldsymbol{V}}_k^{\mathrm{T}}\widehat{\boldsymbol{x}}]\boldsymbol{e}^{\mathrm{T}}, \\ \dot{\widehat{\boldsymbol{V}}}_k = -u_k\Gamma_{\mathrm{V},k}\boldsymbol{x}\boldsymbol{e}^{\mathrm{T}}\widehat{\boldsymbol{W}}_k\widehat{\boldsymbol{S}}'_k. \end{cases} \tag{56}$$

Remark 4. As introduced in Section 1, reinforcement learning control (RLC) methods give iterative algorithms which can be further classified into policy iteration (PI) based RLCs and value iteration (VI) based RLCs. For VI-based RLCs, it is proven that the iterative performance index function is a nondecreasing sequence and bounded, which converges to the optimal performance index when the iteration index increases to infinity. However, the VI-based RLC has a very low computation efficiency, and the closed-loop stability cannot be always provided. Actually, VI-based RLCs are still implemented offline, which seriously limits their practical applications [38]. Thus, for online implementation and satisfactory closed-loop stability, PI-based RLCs should be designed for the operational performance optimization of energy systems. Actually, Algorithm 1 is a typical PI-based RLC, which needs to solve algebraic Riccati equation (52) on time. Since there are many mature numerical programs that can be utilized for solve Riccati equation, it is convenient to apply Algorithm 1 for real-time optimization. The estimations of state-vector $\widehat{\boldsymbol{x}}$ and MLP weighting matrices $\widehat{\boldsymbol{W}}_k$ and$\widehat{\boldsymbol{V}}_k(k = 1, \ldots, m)$ that determines $\widehat{\boldsymbol{A}}_{\mathrm{NN}}$ and $\widehat{\boldsymbol{B}}_{\mathrm{NN}}$in algebraic Riccati equation (52) are given by learning law (56). In the practical implementation, the program of this MLP-based RLC is composed by the program for solving Riccati equation (52) and the program for realizing MLP-based identifier (56), which can be easily deployed on the widely applied digital control system (DCS) platforms.

Remark 5. Both matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ are strict positive-definite, and can be chosen by the operators of energy systems, where $\boldsymbol{Q}$ denotes the weight of the deviation of state-vector $\boldsymbol{x}$ in the performance index, and $\boldsymbol{R}$ denotes the weight of the cost for control action. Actually, both the changes of $\boldsymbol{Q}$ and $\boldsymbol{R}$ can influence the weight of state-vector deviation and control cost in the performance index, and usually we can change $\boldsymbol{Q}$ or $\boldsymbol{R}$ to modify the tradeoff between the deviation of state response and the cost of control input. Moreover, algebraic Riccati equation (52) is solved by the Schur vector approach (SVA) given in [53] and [54]. It is shown in [53] that SVA provides substantially efficient, useful and reliable technique for numerically solving algebraic Riccati equations. Since matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ are assumed to be strict positive-definite, the matrix pair $(\widehat{\boldsymbol{A}}_{\mathrm{NN}}, \boldsymbol{R}^{\frac{1}{2}}\widehat{\boldsymbol{B}}_{\mathrm{NN}})$ is controllable, and $(\boldsymbol{Q}^{\frac{1}{2}}, \widehat{\boldsymbol{A}}_{NN})$ is observable, i.e. both the Gram matrix criteria of controllability and observability are satisfied. Therefore, the conditions of applying SVA given in [53] can be satisfied, and SVA can provide an efficient and reliable solution of algebraic Riccati equation (52).

5.2. Closed-Loop stability analysis

The results in Sections 3 and 4 show that MLP-based RLC (51) can approximate the optimal control (9) effectively. However, the engineering implementation needs not only optimality but also stability. The following Proposition 3 shows that the UUB closed-loop stability can be well guaranteed by the MLP-based RLC (51), which is the third main result of this paper.

Proposition 3. *The closed-loop constituted by nonlinear dissipative system (1) and MLP-based RLC determined by equations (51)–(56) is UUB stable.*

Proof. Choose the Lyapunov function of the closed-loop system as

$$V(\widehat{\boldsymbol{x}}, \boldsymbol{e}, \widetilde{\boldsymbol{W}}_k, \widetilde{\boldsymbol{V}}_k) = \widehat{\boldsymbol{x}}^{\mathrm{T}}\widehat{\boldsymbol{P}}\widehat{\boldsymbol{x}} + \frac{8\lambda_{\max}(\boldsymbol{H}\boldsymbol{P}\boldsymbol{Q}^{-1}\boldsymbol{P}\boldsymbol{H})}{\lambda_{\min}(2\boldsymbol{\Omega} + \boldsymbol{H})}V_{\mathrm{E}}(\boldsymbol{e}, \widetilde{\boldsymbol{W}}_k, \widetilde{\boldsymbol{V}}_k), \tag{57}$$

where matrix $\widehat{\boldsymbol{P}}$ is the solution of Riccati equation (52), and V_{E} is defined in (45).

Differentiate V along the trajectory of closed-loop dynamics given by (1) and (51),

$$\begin{aligned} \dot{V}(\widehat{\boldsymbol{x}}, \boldsymbol{e}, \widetilde{\boldsymbol{W}}_k, \widetilde{\boldsymbol{V}}_k) &= 2\widehat{\boldsymbol{x}}^{\mathrm{T}}\widehat{\boldsymbol{P}}\dot{\widehat{\boldsymbol{x}}} + \frac{8\lambda_{\max}(\boldsymbol{H}\boldsymbol{P}\boldsymbol{Q}^{-1}\boldsymbol{P}\boldsymbol{H})}{\lambda_{\min}(2\boldsymbol{\Omega} + \boldsymbol{H})}\{\boldsymbol{e}^{\mathrm{T}}\dot{\boldsymbol{e}} \\ &\quad + \sum_{k=0}^{m}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_k^{\mathrm{T}}\Gamma_{\mathrm{W},k}^{-1}\dot{\widehat{\boldsymbol{W}}}_k) + \mathrm{tr}(\widetilde{\boldsymbol{V}}_k^{\mathrm{T}}\Gamma_{\mathrm{V},k}^{-1}\dot{\widehat{\boldsymbol{V}}}_k)]\} \\ &\leqslant \widehat{\boldsymbol{x}}^{\mathrm{T}}\{\widehat{\boldsymbol{P}}\,[\widehat{\boldsymbol{A}}_{\mathrm{NN}}(\widehat{\boldsymbol{x}}) - \boldsymbol{\Omega}] + [\widehat{\boldsymbol{A}}_{\mathrm{NN}}(\widehat{\boldsymbol{x}}) - \boldsymbol{\Omega}]^{\mathrm{T}}\widehat{\boldsymbol{P}} - \widehat{\boldsymbol{P}}\widehat{\boldsymbol{B}}_{\mathrm{NN}}(\widehat{\boldsymbol{x}})\boldsymbol{R}^{-1}\boldsymbol{B}_{\mathrm{NN}}^{\mathrm{T}}(\widehat{\boldsymbol{x}})\widehat{\boldsymbol{P}}\}\widehat{\boldsymbol{x}} \\ &\quad - 2\widehat{\boldsymbol{x}}^{\mathrm{T}}\widehat{\boldsymbol{P}}\boldsymbol{H}\boldsymbol{e} \\ &\quad - 4\lambda_{\max}(\boldsymbol{H}\widehat{\boldsymbol{P}}\boldsymbol{Q}^{-1}\widehat{\boldsymbol{P}}\boldsymbol{H})\lambda_{\min}^{-1}(2\boldsymbol{\Omega} + \boldsymbol{H})[\boldsymbol{e}^{\mathrm{T}}(2\boldsymbol{\Omega} + \boldsymbol{H})\boldsymbol{e} - \boldsymbol{d}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{d}_{\mathrm{A}}] \\ &= -\widehat{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{Q}\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}^{\mathrm{T}}\widehat{\boldsymbol{P}}\boldsymbol{H}\boldsymbol{e} - 4\lambda_{\max}(\boldsymbol{H}\widehat{\boldsymbol{P}}\boldsymbol{Q}^{-1}\widehat{\boldsymbol{P}}\boldsymbol{H})[\|\boldsymbol{e}\|_2^2 \\ &\quad - \lambda_{\min}^{-1}(2\boldsymbol{\Omega} + \boldsymbol{H})\lambda_{\min}^{-1}(\boldsymbol{H})\|\boldsymbol{d}_{\mathrm{A}}\|_2^2] \\ &\leqslant -\tfrac{1}{2}\widehat{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{Q}\widehat{\boldsymbol{x}} - 2\lambda_{\max}(\boldsymbol{H}\widehat{\boldsymbol{P}}\boldsymbol{Q}^{-1}\widehat{\boldsymbol{P}}\boldsymbol{H})\|\boldsymbol{e}\|_2^2 \\ &\quad + 4\lambda_{\max}(\boldsymbol{H}\widehat{\boldsymbol{P}}\boldsymbol{Q}^{-1}\widehat{\boldsymbol{P}}\boldsymbol{H})\lambda_{\min}^{-1}(2\boldsymbol{\Omega} + \boldsymbol{H})\lambda_{\min}^{-1}(\boldsymbol{H})\|\boldsymbol{d}_{\mathrm{A}}\|_2^2, \end{aligned} \tag{58}$$

from which it can be seen that the closed-loop state-vector $\boldsymbol{z} = [\widehat{\boldsymbol{x}}^{\mathrm{T}} \;\; \boldsymbol{e}^{\mathrm{T}}]^{\mathrm{T}}$ finally enters to the bounded set given by

$$\Theta = \{\boldsymbol{z} \in \mathrm{R}^{2n} \mid \boldsymbol{z}^{\mathrm{T}}\Phi\boldsymbol{z} \leqslant 4\lambda_{\max}(\boldsymbol{H}\widehat{\boldsymbol{P}}\boldsymbol{Q}^{-1}\widehat{\boldsymbol{P}}\boldsymbol{H})\lambda_{\min}^{-1}(2\boldsymbol{\Omega} + \boldsymbol{H})\lambda_{\min}^{-1}(\boldsymbol{H})\|\boldsymbol{d}_{\mathrm{A}}\|_2^2\}, \tag{59}$$

where

$$\Phi = \frac{1}{2}\begin{bmatrix} \boldsymbol{Q} & \boldsymbol{O} \\ \boldsymbol{O} & 2\lambda_{\max}(\boldsymbol{H}\widehat{\boldsymbol{P}}\boldsymbol{Q}^{-1}\widehat{\boldsymbol{P}}\boldsymbol{H})\boldsymbol{I}_n \end{bmatrix}. \tag{60}$$

This is to say that the closed-loop is UUB, and thus the proof of Proposition 3 is completed.

6. Application to the optimized NSSS thermal power control

In this section, the MLP-based RLC given in the above section is applied to optimize the thermal power response of the NSSS of HTR-PM plant [55,56]. The NSSS is composed of a modular high temperature gas-cooled reactor (MHTGR), a helical-coil once-through steam-generator (OTSG) and connecting pipelines. The simplified diagram of the MHTGR-based NSSS is given in Fig. 3, where the MHTGR uses helium as coolant and graphite as the moderator as well as structural material, and live steam generated by the OTSG is superheated [56,57]. Due to the low power density, full-power-range negative temperature feedback effect and high surface-to-volume ratio, the MHTGR has inherent nuclear safety feature, and is seen as one of the candidates for the next generation of nuclear plant [56]. Further, the MHTGR-based nuclear plant can take the multimodular scheme, namely, the live steam from a number of NSSS modules are combined together to drive a common steam turbine, which can simultaneously meet the requirements of safety and economy. The thermal power control of MHTGR-based NSSS is important in balancing the thermal power supply and demand in a nuclear plant. However, the local PID controllers, which regulate

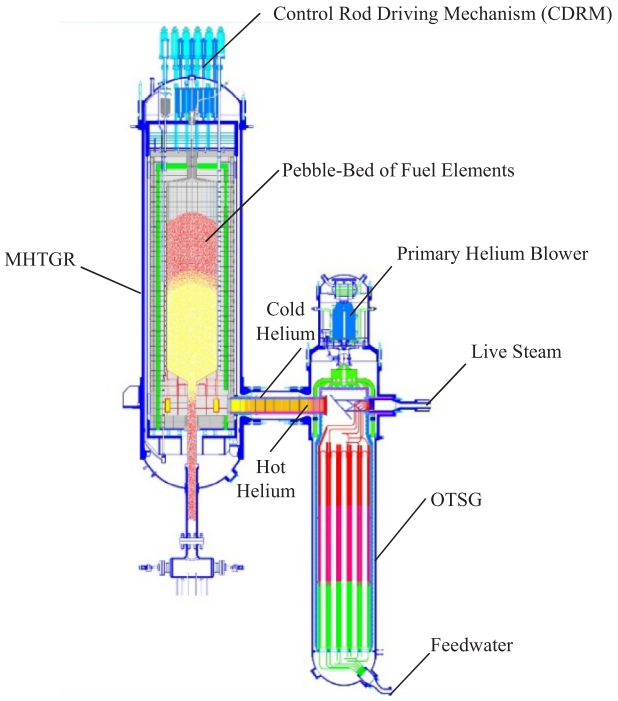

Fig. 3. Schematic diagram of the NSSS module of HTR-PM Plant.

process variables such as the neutron flux as well as coolant temperatures, pressures and flowrates, cannot directly control the NSSS thermal power. Hence, the NSSS thermal power should be regulated at the supervisory control level, which gives the motivation to apply the MLP-based RLC composed of (51)–(56) for the supervisory control of NSSS thermal power.

6.1. Implementation of MLP-based RLC

The closed-loop structure of MLP-based RLC for NSSS thermal power response optimization is given in Fig. 4, where n_r is the normalized neutron flux, T_{cout} is the hot helium temperature at reactor outlet, G_h is the primary helium flowrate, T_{st} is the live steam temperature, G_{fw} is the OTSG feedwater flowrate, P_{th} is the NSSS thermal power, scalars n_{r0}, T_{cout0}, G_{h0}, T_{st0}, G_{fw0} and P_{th0} are the setpoints of n_r, T_{cout}, G_h, T_{st}, G_{fw} and P_{th} respectively, and δn_{r0} is the revision to the setpoint of normalized neutron flux. As shown in Fig. 4, the model-free adaptive control (MAC) presented in [30] is adopted as the inner-loop local controller, namely, to regulate the key process variables of the MHTGR-based NSSS module of HTR-PM plant including the neutron flux, primary helium flowrate and hot helium temperature at reactor outlet as well as secondary feedwater flowrate and live steam temperature so as to keep their measurements in a bounded neighborhood of their setpoints. The supervisory controller located in the outer loop adopts the MLP-based RLC which generates proper normalized neutron flux setpoint revision δn_{r0} so as to guarantee the stability of NSSS thermal power deviation δP_{th} through minimizing the performance index given by equations (3) and (4) with $m = 1$,

$$u_1 = \delta n_{r0}, \tag{61}$$

$$x = \left[\int_{t_0}^{t} \delta P_{th}(\tau) d\tau \quad \delta P_{th} \right]^T, \tag{62}$$

where δn_{r0} is the setpoint revision of normalized neutron flux, $\delta P_{th} = P_{th}$-P_{th0}, P_{th} is the NSSS thermal power, P_{th0} is the setpoint of P_{th}, and norm-bounded vector functions $\boldsymbol{f}_0$ and $\boldsymbol{f}_1$ are approximated by MLPs with learning laws given by (43) and (44).

6.2. Simulation results

This simulation is performed on Matlab/Simulink environment, where the dynamical model of HTR-PM plant adopts the one presented in [58]. This model can reflect the main dynamical behavior of the MHTGR-based NSSS, and has been verified through the comparison to the actual response of HTR-10 test reactor [59] that is the prototype of HTR-PM. In this simulation, the parameters of the newly-built MLP-based predictive controller is given as $l = 4$, $\Omega = 0.001\boldsymbol{I}_2$, $\boldsymbol{Q} = \text{diag}([0.001, 1])$, $H = 0.1$, and R is chosen to be different values so as to show its influence to dynamic responses. The parameters of the MAC in the inner loop for regulating key process variables are the same with those given in [30]. The case of NSSS module power maneuvering between 100% and 50% reactor full power (RFP) is considered. Initially, the HTR-PM plant operates at 100% plant full power (PFP), i.e. two NSSS modules runs at 100%RFP. The thermal power setpoint of 1# NSSS module starts to ramp down to 50%RFP at 3000 s with a constant rate of 5%RFP/min, and then starts to ramp back to 100%RFP at 6000 s with the same rate. The following two comparisons are studied in this numerical simulation.

(1) Comparison between the cases with and without MLP-based RLC

The comparison result is shown in Figs. 5 and 6 respectively, where Fig. 5 gives the responses of normalized neutron flux, hot helium temperature at reactor outlet, live steam temperature of OTSG and NSSS thermal power, and Fig. 6 shows the response of thermal power error δP_{th} and control input δn_{r0} that given by the MLP-based RLC. Here, R is chosen as $R = 2e4$ for the case with the MLP-based RLC. In Figs. 5 and 6, the red solid lines and the black dot lines correspond to the responses of process variables and control input with and without the optimization provided by the MLP-based RLC respectively. From

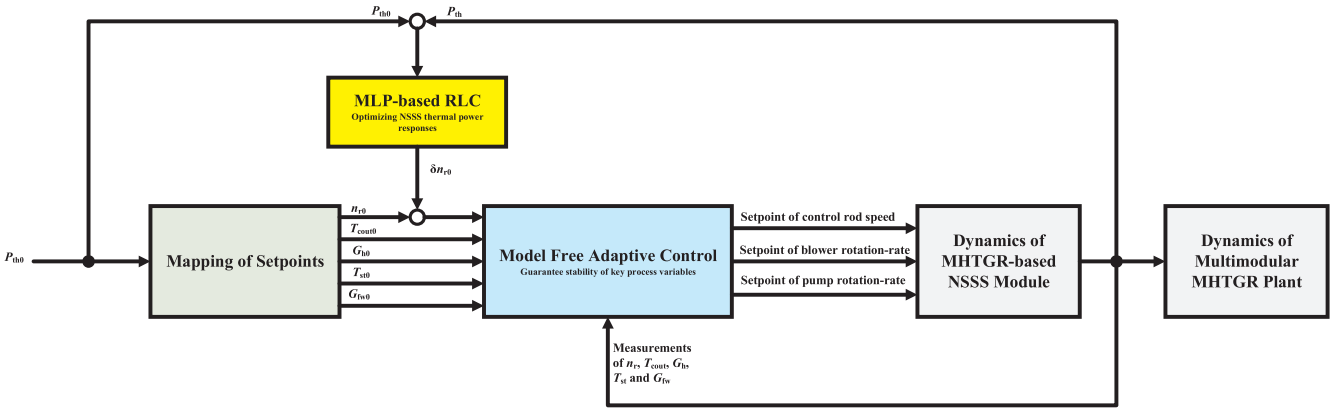

Fig. 4. Closed-loop structure of MLP-based RLC for thermal power response optimization.

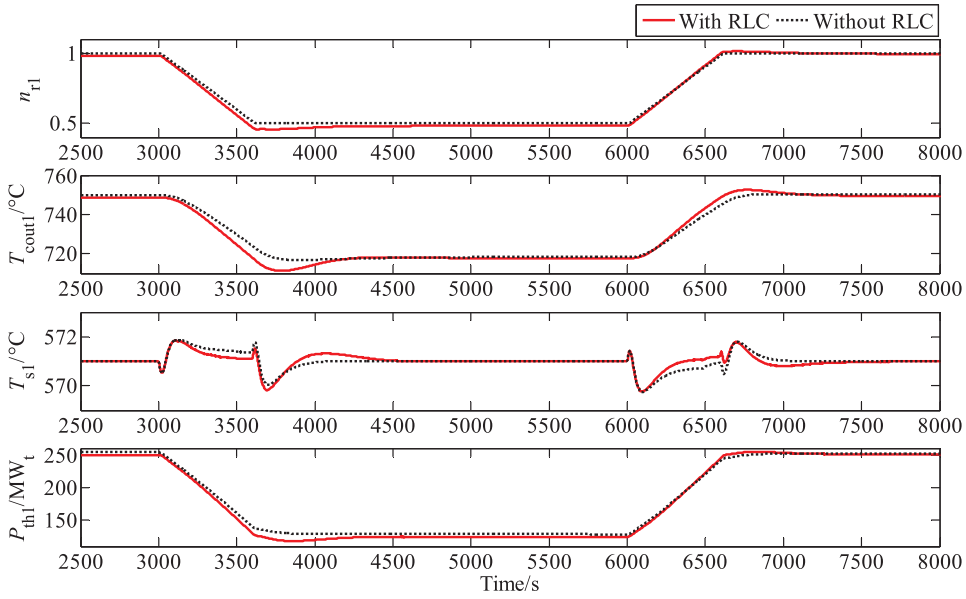

Fig. 5. Responses of 1#NSSS process variables in the comparison between the cases with and without reinforcement learning control, n_{r1}: normalized nuclear power, T_{cout1}: reactor outlet helium temperature, T_{s1}: live steam temperature, P_{th1}: NSSS thermal power.

Fig. 6, it can be seen that the steady error of NSSS thermal power is eliminated while the magnitude of thermal power overshoot is also decreased. From Fig. 5, due to the neutron flux is revised by the MLP-based RLC so as to optimize the thermal power response, the overshoots of both neutron flux and hot helium temperature become a little larger and are still satisfactory.

(2) Influence of different values of R

The comparison among the results corresponding to R = 5e4, 2e4, 1e4 and 5e3 is shown in Figs. 7 and 8 respectively, where Fig. 7 gives the responses of neutron flux, hot helium temperature at the reactor outlet, live steam temperature and NSSS thermal power, and Fig. 8 gives the responses of both thermal power error δP_{th} and control input δn_{r0}. The red solid lines, blue dash-dot lines, black dot lines and green dash lines correspond to value of R equals 5e4, 2e4, 1e4 and 5e3 respectively. It can be seen from Figs. 7 and 8 that, R is larger, the oscillations of neutron flux n_{r1}, hot helium temperature T_{cout1} and steam temperature T_{s1} are all smaller while both the overshoot and transition time of NSSS thermal power error δP_{th} being larger. Further, from Fig. 8, a larger R leads to a larger control input δn_{r0} and a smaller steady error of NSSS thermal, however, the oscillation of thermal power error is stronger.

6.3. Discussions

The power ramping of a NSSS module is induced by the ramping of its thermal power setpoint that is actively triggered by the reactor operator. The setpoint of NSSS thermal power determines the setpoints of NSSS key process variables including the reactor neutron flux as well as the temperatures and flowrates of primary helium flow and secondary water/steam flow. During the power ramping of a NSSS module, the ramping of NSSS thermal power setpoint enlarges the errors between measurements and setpoints of NSSS key process variables, which drives the MAC-based local controller to adjust the setpoints of control rod speed and rotation-rates of both the helium blower and feedwater pump so as to suppress these control errors. As the NSSS thermal power is not directly regulated by the local controller, it is necessary to optimize the thermal power response so as to obtain a better balance between the heat supply and demand, which induce applying the MLP-based RLC given in this paper for thermal power transient optimization.

The difference of dynamic responses in the cases with and without RLC is given by the optimization action provided by the MLP-based RLC proposed in this paper. Although the MAC-based local control strategy

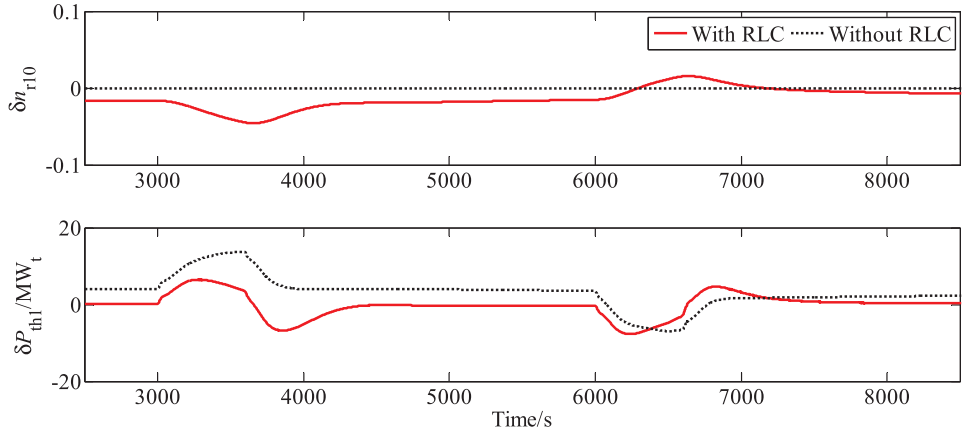

Fig. 6. Responses of control input and thermal power error of 1# NSSS in the comparison between the cases with and without reinforcement learning control, δn_{r01}: revision to the setpoint of normalized nuclear power, δP_{th1}: thermal power error.

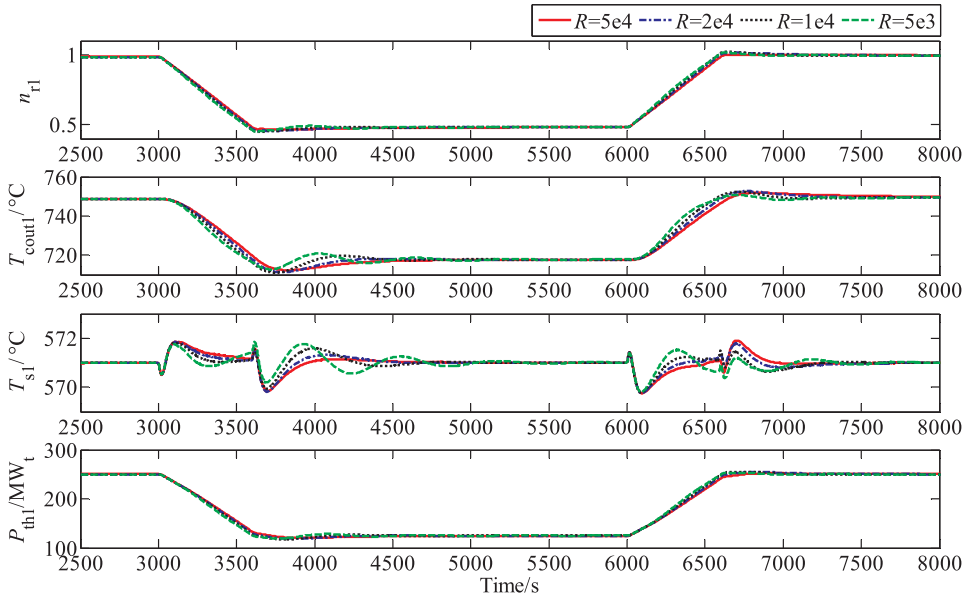

Fig. 7. Responses of 1#NSSS process variables with different values of R, n_{r1}: normalized nuclear power, T_{cout1}: reactor outlet helium temperature, T_{s1}: live steam temperature, P_{th1}: NSSS thermal power.

guarantees the stability of neutron flux as well as the temperatures and flowrates of primary helium and secondary superheated steam, the NSSS thermal power is NOT directly controlled locally. From Fig. 4, the MLP-based RLC serves in the outer loop as a supervisory controller which provides the stability of NSSS thermal power by revising the setpoint of neutron flux. Based upon equations (3) and (4), since the performance index is partially given by penalty from the deviation of thermal power, and the control action is calculated for minimizing the performance index, the deviation of NSSS thermal power must be suppressed to some extent. Here, the process of the MLP-based RLC generating a control action is given as follows: First, the RLC first identifies the dynamics from the setpoint revision of normalized neutron flux to the NSSS thermal power online by MLP-based observer (33) as well as learning laws (43) and (44). Second, the RLC generates the setpoint revision of normalized neutron flux by (51) to suppress the thermal power error, where system input matrix $\widehat{\boldsymbol{B}}_{NN}$ is given by the dynamical identification, and gain matrix $\widehat{\boldsymbol{P}}$ is obtained from solving algebraic Riccati equation (52). From Figs. 5 and 6, the steady error of NSSS thermal power is cleared, and the magnitude of overshoot is reduced effectively. Meanwhile, the deterioration in the responses of reactor neutron flux, hot helium temperature at reactor outlet and OTSG live steam temperature is quite limited and acceptable. This shows the feasibility of applying this newly-built MLP-based RLC for NSSS thermal power optimization.

Moreover, from Figs. 7 and 8, the value of R can deeply influence the control performance. If R is larger, the weight of the cost for adding a control action in utility function r defined in (4) is larger, and the minimization of performance index (3) leads to a smaller magnitude of control input δn_{r0}. Actually, from equation (51), the feedback gain becomes smaller if R is larger. Due to the control input is weakened, the transient response of NSSS thermal power deviation δP_{th1} cannot be enhanced to a satisfactory level. As we can see from Fig. 8, both the overshoot and steady error of δP_{th1} are larger if R is larger. Similarly, if R is smaller, the weight of control cost in utility function r is smaller, and the weight of NSSS thermal power deviation becomes larger. The minimization of the performance index results in a larger δn_{r0} so as to decrease NSSS thermal power deviation δP_{th1}. From Fig. 8, as R becomes smaller, the magnitude control input δn_{r0} is larger, and both the overshoot and steady error of δP_{th1} are smaller. Due to the fact that a lower deviation of δP_{th1} should be guaranteed by a larger magnitude of control input δn_{r0}, there exists a reverse phenomenon in the varying trend of δn_{r0} and δP_{th1} when R becomes larger or smaller. Further from Figs. 7 and 8, if the value of R is too small, then there exists the oscillations of the process variables such as the live steam temperature of

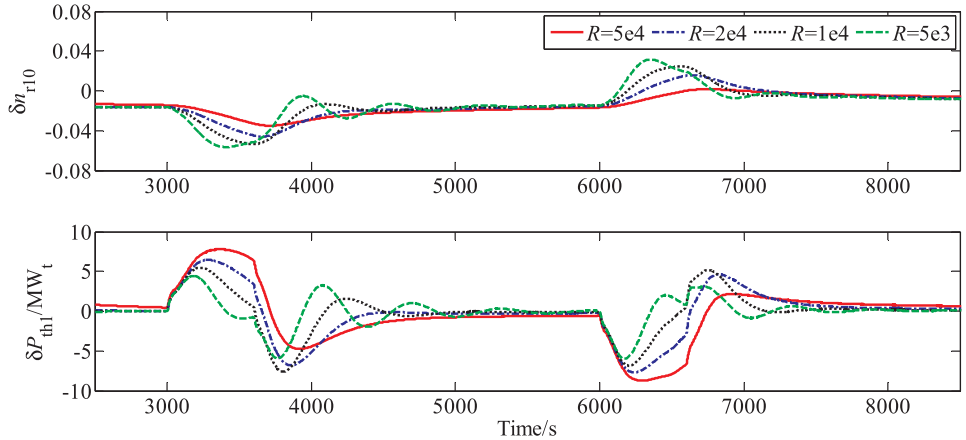

Fig. 8. Responses of control input and thermal power error of 1# NSSS with different values of R, δn_{r01}: revision to the setpoint of normalized nuclear power, δP_{th1}: thermal power error.

OTSG. Hence, in the practical engineering, R should be carefully chosen so as to obtain a good tradeoff between optimizing thermal power response and avoiding oscillation in the responses. Finally, the responses in Figs. 5 and 7 verify the result of UUB closed-loop stability given by Proposition 3.

In a summary, the numerical simulation results verify the results given by Propositions 1–3 about the optimization capability of the newly-built MLP-based RLC as well as the rationality of the closed-loop stability analysis. Further, from (51)–(56), since Algorithm 1 is composed of several ordinary differential and algebraic equations, it is easy to be implemented on those widely-used digital control system platforms.

7. Conclusions

Supervisory control of energy systems, which steers the evolution of the interconnection between energy processes and their local controllers through revising the setpoints of local controllers, is crucial in optimizing the performance indices determined by the factors such as thermal efficiency, steady errors and overshoots. Due to the ability of reinforcement learning control (RLC) in online approximating the optimal control law corresponding to a user-defined utility function, and because of the dissipation feature of the nonlinear system coupled by energy processes and their local controllers, a RLC composed of a multilayer perception (MLP)-based state-observer and an approximated optimal controller is newly proposed. The MLP-based observer provides an identification that asymptotically converges to a bounded neighborhood of the original system dynamics. After giving the linear representation of this MLP-based state-observer, this approximated optimal controller is given by the solution of the Riccati equation corresponding to the linear representation. It has been further shown from Lyapunov direct method that the closed-loop is UUB stable. Finally, the MLP-based RLC is applied to the optimization of thermal power response for a high temperature gas-cooled reactor based nuclear steam supply system (NSSS), and simulation results show not only the feasibility and satisfactory performance but also the influence of the controller parameters to closed-loop responses.

CRediT authorship contribution statement

Zhe Dong: Conceptualization, Methodology, Software, Writing - original draft. **Xiaojin Huang:** Writing - review & editing. **Yujie Dong:** Supervision. **Zuoyi Zhang:** Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is jointly supported by National S&T Major Project of China, China (Grants No. ZX06906 and ZX06902), Natural Science Foundation of China (NSFC), China (Grant No. 61773228) as well as Opening Fund of State Key Laboratory of Nuclear Power Safety Monitoring Technology and Equipment (Grant No. K-A2018.418), China. The authors would like to thank the anonymous reviewers and editors for the valuable comments and suggestions.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apenergy.2019.114193.

References

[1] Sangi R, Müller D. Application of the second law of thermodynamics to control: a review. Energy 2019;174:938–53.
[2] Nguyen A, Lauber J, Dambrine M. Optimal control based algorithms for energy management of automotive power systems with battery/supercapacitor storage devices. Energy Convers Manage 2014;87:410–20.
[3] Costanza V, Rivadeneira PS. Optimal supervisory control of steam generator in parallel. Energy 2015;93:1819–31.
[4] Chassin DP, Behboodi S, Shi Y, Djilali N. H_2-optimal transactive control of electric power regulation from fast-acting demand response in the presence of high renewables. Appl Energy 2017;205:304–15.
[5] Hemi H, Chouili J, Cheriti A. Combination of Markov chain and optimal control solved by Pontryagin's minimum principle for a fuel cell/supercapacitor vehicle. Energy Convers Manage 2015;91:387–93.
[6] Hou C, Ouyang M, Xu L, Wang H. Approximate Pontryagin's minimum principle applied to the energy management of plug-in hybrid electric vehicles. Appl Energy 2014;115:174–89.
[7] Onori S, Tribioli L. Adaptive Pontryagin's minimum principle supervisory controller design for the plug-in hybrid GM Chevrolet Volt. Appl Energy 2015;147:224–34.
[8] Zhang S, Xiong R, Zhang C. Pontryagin's minimum principle-based power management of a dual-motor-driven electric bus. Appl Energy 2015;159:370–80.
[9] Jain N, Alleyne A. Exergy-based optimal control of a vapor compression system. Energy Convers Manage 2015;92:353–65.
[10] Oravec J, Bakošová M, Trafczynski M, Vasičkaninová A, Mészáros A, Markowski M. Robust model predictive control and PID control of shell-and-tube heat exchangers. Energy 2018;159:1–10.
[11] Cox SJ, Kim D, Cho H, Mago P. Real time optimal control of district cooling system with thermal energy storage using neural networks. Appl Energy 2019;238:466–80.
[12] Wu X, Wang M, Shen J, Lawal A, Lee KY. Reinforced coordinated control of coal-fired power plant retrofitted with solvent based CO_2 capture using model predictive controls. Appl Energy 2019;238:495–515.
[13] Finck C, Li R, Zeiler W. Economic model predictive control for demand flexibility of a residential building. Energy 2019;176:365–79.
[14] Drgoňa J, Picard D, Kvasnica M, Helsen L. Approximate model predictive building control via machine learning. Appl Energy 2018;218:199–216.
[15] Mirakhorli A, Dong B. Model predictive control for building loads connected with a residential distribution grid. Appl Energy 2018;230:627–42.
[16] Gholamibozanjani G, Tarragona J, de Gracia A, Ferández C, Cabeza LF, Farid MM. Model predictive control strategy applied to different types of building for space heating. Appl Energy 2018;231:959–71.
[17] Blum DH, Arendt K, Rivalin L, Piette MA, Wetter M, Veje CT. Practical factors of envelope model setup and their effects on the performance of model predictive control for building heating, ventilating, and air conditioning systems. Appl Energy 2019;236:410–25.
[18] Bianchini G, Casini M, Pepe D, Vicino A, Zanvettor GG. An integrated model predictive control approach for optimal HVAC and energy storage operation in large-scale buildings. Appl Energy 2019;240:327–40.
[19] Tang R, Wang S. Model predictive control for thermal energy storage and thermal comfort optimization of building demand response in smart grids. Appl Energy 2019;242:873–82.
[20] Joe J, Karava P. A model predictive control strategy to optimize the performance of radiant floor heating and cooling systems in office buildings. Appl Energy 2019;245:65–77.
[21] Hu J, Xu Y, Cheng KW, Guerrero JM. A model predictive control strategy of PV-Battery microgrid under variable power generations and load conditions. Appl Energy 2018;221:195–203.
[22] Romero-Quete D, Garcia JR. An affine arithmetic-model predictive control approach for optimal economic dispatch of combined heat and power microgrids. Appl Energy 2019;242:1436–47.
[23] Lv C, Yu H, Li P, Wang C, Xu X, Li S, et al. Model predictive control based robust scheduling of community integrated energy system with operational flexibility. Appl Energy 2019;243:250–65.
[24] Xie S, Hu X, Xin Z, Brighton J. Pontryagin's minimum principle based model predictive control of energy management for a plug-in hybrid electric bus. Appl Energy 2019;236:893–905.
[25] Haseltalab A, Negenborn RR. Model predictive maneuvering control and energy management for all-electric autonomous ships. Appl Energy 2019;251:113308.
[26] Na MG, Shin SH, Kim WC. A model predictive controller for nuclear reactor power. J Kor Nucl Soc 2003;35:399–411.
[27] Na MG, Hwang IJ, Lee YJ. Design of a fuzzy model predictive power controller for pressurized water reactors. IEEE Trans Nucl Sci 2006;53:1504–14.
[28] Etchepareborda A, Lolich J. Research reactor power controller design using an output feed-back nonlinear receding horizon control method. Nucl Eng Des 2007;237:268–76.
[29] Eliasi H, Menhaj MB, Davilu H. Robust nonlinear model predictive control for a PWR nu-clear power plant. Prog Nucl Energy 2012;54:177–85.
[30] Dong Z, Pan Y, Zhang Z, Dong Y, Huang X. Model-free adaptive control law for nuclear superheated-steam supply systems. Energy 2017;135:53–67.
[31] Jiang D, Dong Z. Practical dynamic matrix control of MHTGR-based nuclear steam supply systems. Energy 2019;185:695–707.
[32] Dong Z, Zhang Z, Dong Y, Huang X. Multi-layer perception based model predictive control for the thermal power of nuclear superheated-steam supply systems. Energy 2018;151:11–125.
[33] Yang L, Nagy Z, Goffin P, Schlueter A. Reinforcement learning for optimal control of

low exergy buildings. Appl Energy 2015;156:577–86.
[34] Lewis FL, Vrabie D, Vamvoudakis KG. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. IEEE Control System Mag 2012;32:76–105.
[35] Lewis FL, Vamvoudakis KG. Reinforcement learning for partially observable dynamic processes: adaptive dynamic programing using measured output data. IEEE Trans Syst Man Cybernet Part B: Cybernet 2011;41:14–25.
[36] Kiumarsi B, Vamvoudakis KG, Modares H, Lewis FL. Optimal and autonomous control using reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst 2018;29:2042–62.
[37] Al-Tamimi A, Lewis FL, Abu-Khalaf M. Discrete-time nonlinear HJB solution using ap-proximate dynamic programming: convergence proof. IEEE Trans Syst Man Cybernet Part B: Cybernet 2008;38:943–9.
[38] Liu D, Wei Q. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. IEEE Trans Neural Netw Learn Syst 2014;25:621–34.
[39] Liu D, Wang D, Li H. Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach. IEEE Trans Neural Netw Learn Syst 2014;25:418–28.
[40] Liu D, Yang X, Wang D, Wei Q. Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints. IEEE Trans Cybern 2015;45:1372–85.
[41] Sprangers O, Babuška R, Nageshrao SP, Lopes GAD. Reinforcement learning for port-Hamiltonian systems. IEEE Trans Cybern 2015;45:1003–13.
[42] Vamvoudakis KG, Lewis FL. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. Automatica 2010;46:878–88.
[43] Bhasin S, Kamalapurkar R, Johnson M, Vamvoudakis KG, Lewis FL, Dixon WE. A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. Automatica 2013;49:82–92.
[44] Kamalapurka B, Walters P, Dixon WE. Model-based reinforcement learning for approximate optimal regulation. Automatica 2016;64:94–104.
[45] Kamalapurkar R, Andrews L, Walers P, Dixon WE. Model-based reinforcement learning for infinite-horizon approximate optimal tracking. IEEE Trans Neural Networks Learn Syst 2017;28:753–8.
[46] Valasek J, Doebbler J, Tandale MD, Meade AJ. Improved adaptive-reinforcement learning control for morphing unmanned air vehicles. IEEE Trans Syst Man Cybernet Part B: Cybernet 2008;38:1014–20.
[47] Yan Z, Xu Y. Data-driven load frequency control for stochastic power systems: a deep reinforcement learning method with continuous action search. IEEE Trans Power Syst 2019;34:1653–6.
[48] El-Tantawy S, Abdulhai B, Abdelgawad H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto. IEEE Trans Intell Transp Syst 2013;14:1140–50.
[49] Jiang Y, Fan J, Chai T, Li J, Lewis FL. Data-driven flotation industrial process operational optimal control based on reinforcement learning. IEEE Trans Ind Inf 2018;14:1974–89.
[50] Modares H, Ranatunga I, Lewis FL, Popa DO. Optimized assistive human-robot interaction using reinforcement learning. IEEE Trans Cybern 2016;46:655–67.
[51] Demirel B, Ramaswamy A, Quevedo DE, Karl H. DeepCAS: a deep reinforcement learning algorithms for control-aware scheduling. IEEE Control Syst Lett 2018;2:737–42.
[52] Ge SS, Huang CC, Lee TH, Zhang T. Stable adaptive neural network control. The Netherlands: Kluwer Academic Publishers; 2002.
[53] Laub AJ. A Schur method for solving algebraic Riccati equations. IEEE Trans Autom Control 1979;24:913–21.
[54] Arnold III WF, Laub AJ. Generalized eigenproblem algorithms and software for algebraic Riccati equations. Proc IEEE 1984;72:1746–54.
[55] Zhang Z, Sun Y. Economic potential of modular reactor nuclear power plants based on the Chinese HTR-PM project. Nucl Eng Des 2007;237:2265–74.
[56] Zhang Z, Dong Y, Li F, et al. The shandong shidao bay 200 MWe high-temperature-gas-cooled reactor pebble-bed module (HTR-PM) demonstration power plant: An engineering and technological innovation. Engineering 2016;2:112–8.
[57] Lohnert GH. Technical design features and essential safety-related properties of the HTR-MODULE. Nucl Eng Des 1990;121:259–75.
[58] Dong Z, Pan Y, Song M, Huang X, Dong Y, Zhang Z. Dynamic modeling and control characteristics of the two-modular HTR-PM nuclear plant. Sci Technol Nucl Instal 2017;2017:6298037.
[59] Hu S, Liang X, Wei L. Commissioning and operation experience and safety experiments on HTR-10. The 3rd International Topical Meeting on High Temperature Reactor. Technology 2006:D00000052.